A 65 nm, 850 MHz, 256 kbit, 4.3 pJ/access, ultra low leakage power memory using dynamic cell stability and a dual swing data link

Bram Rooseleer[1], Stefan Cosemans[1,2] and Wim Dehaene[1,2]
[1]ESAT-MICAS, K.U.Leuven, [2]IMEC
Leuven, Belgium
Email: bram.rooseleer@esat.kuleuven.be

***Abstract*—This paper presents a 65 nm, 256 kbit SRAM memory which achieves both ultra low leakage power and very low active energy consumption at a speed of 850 MHz. Used techniques include divided word and bitlines, local write sense amplifiers, dynamic cell stability and a distributed decoder. In addition, three novel techniques are proposed which decrease power consumption even further. High threshold voltage cells reduce leakage and improve stability. Dual swing signalling on the global bitlines reduces energy without compromising robustness. The decoder uses a new type of dynamic gate to increase speed.**

The design was fabricated in a low power 65 nm CMOS process. Measured performance for this 256 kbit SRAM with 32 bit wordlength is 4.3 pJ per access and 25.2 μW leakage power at a speed of 850 MHz.

I. Introduction

Mobile applications such as smart phones and sensor networks are only possible when power consumption is low enough to ensure a sufficiently long stand-alone time. Increasing performance demands make it difficult for designers to meet the specifications. The digital part of these system-on-chips consists for an ever increasing part out of memories, making those memories an important factor in the total energy consumption [1], [2]. Systems with relatively low duty cycles lead to a focus shift from active to leakage power. In this paper, a $256\,\mathrm{kbit}$ SRAM memory is designed for a speed of $800\,\mathrm{MHz}$. The main focus of this design is low power (both leakage and active) at relatively high speed. A summary of this design is given in Table I.

Section II describes the global structure of the memory. Section III describes the distributed decoder which was used to save energy. Section IV deals with cell design and the techniques that were used to ensure cell stability while maintaining low leakage and without compromising speed. In section V, a new technique, using two different low swing voltages on the global bitlines, is introduced to significantly reduce the impact of transistor mismatch on memory performance. At the same time, robustness can be improved. Section VI discusses the measurement results of the fabricated test chip.

TABLE I
Memory summary.

Memory size	256 kbit
Wordlength	32 bit
Cell type	6T-SRAM cell ($L = 60\,\mathrm{nm}$, $W = 120\,\mathrm{nm}$)
Technology	65 nm low standby power triple-V_T CMOS
Retention mode leakage	25.2 μW
Energy	4.3 pJ/access
Maximal speed	850 MHz
Area	$0.48\,\mathrm{mm}^2$
Known techniques	Divided bitlines Divided wordlines Dynamic cell stability Local write sense amplifier Distributed decoder
Innovations	Dual swing data link on the GBLs High threshold voltage cells Dynamic decoder with merged address latches

II. Memory organisation

To keep energy per access low, it is important to limit unnecessary activity. For a memory, this means that only cells which need to be read should be accessed. To achieve this, wordlines need to be divided on a word-by-word basis, leading to a fully divided wordline architecture. Reducing bitline load has a number of beneficial effects such as an increase in speed and a reduction in energy [3]. In this design, this load is decreased by using divided bitlines. In addition, cell stability (section IV) can be improved with this technique. The combination of divided word and bitlines elegantly results in local memory subblocks.

A. Global structure

Fig. 1 shows the global structure of the memory. Vertically, it is split into two parts to limit the RC-delay on the long horizontal lines. Each part consists of 8 columns (BC) of local blocks (LB). In the ortoghonal direction, the memory is divided in 16 rows (BR) of local blocks. All block columns have 32 pairs of complementary vertical global bitlines (VGBL). The VGBLs are multiplexed (MUX) to one set of 32 pairs of horizontal global bitlines (HGBL), which allows global sense amplifiers (GSA), write drivers and precharge circuitry (GBLD) to be shared between all block columns. A global control circuit (CONT) controls all memory activity.

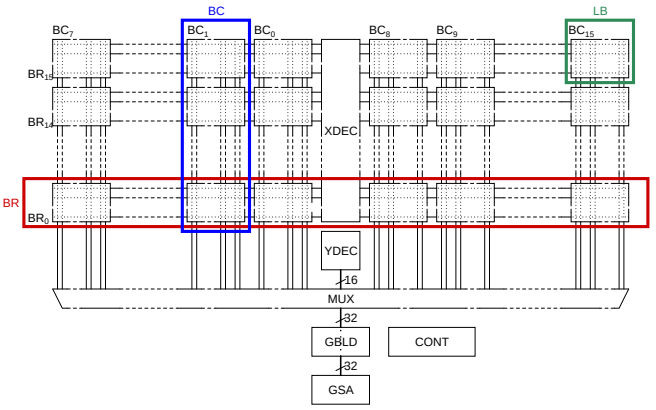


Fig. 1. The global structure of the memory. It is divided in columns and rows of local blocks.

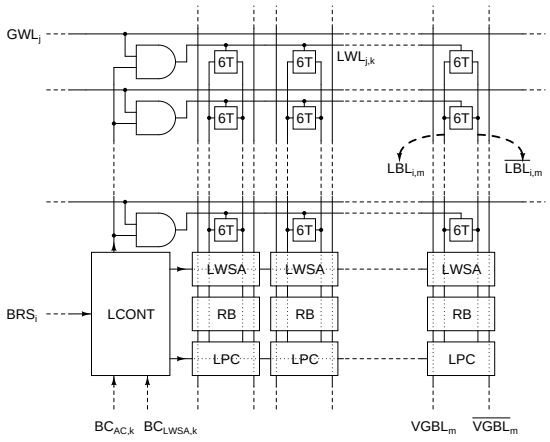


Fig. 2. The local block architecture. The local block contains a matrix of cells, read buffers, local write sense amplifiers, precharge circuitry and control.

B. Local block

As shown in Fig. 2, the local block contains a matrix of 32 words connected with local bitlines (LBL) and some peripheral circuits to read or write these words. Read buffers (RB) connect the local bitlines with the VGBLs. Local write sense amplifiers (LWSA) amplify the low swing signals on the VGBLs to full swing signals on the local bitlines (LBL) during write operations. Additionally, the local block contains a precharge circuit (LPC) and control logic (LCONT).

III. DECODERS

A. Distributed decoder

In order to select a specific local block, a vertically and a horizontally routed signal are needed. For each block row, a horizontally routed signal called block row select (BRS), indicating that one of the local blocks on that row is active, is produced by the X-decoder (XDEC) as shown in Fig. 3. For each block column, two signals $\mathrm{BC_{AC}}$ and $\mathrm{BC_{LWSA}}$ which control local behavior are generated by the Y-decoder (YDEC). $\mathrm{BC_{AC}}$ controls cell access for both read and write operations while $\mathrm{BC_{LWSA}}$ enables the LWSAs. The global wordlines (GWL), shared by all local blocks in the same block row, indicate which specific word within the local block must be activated.

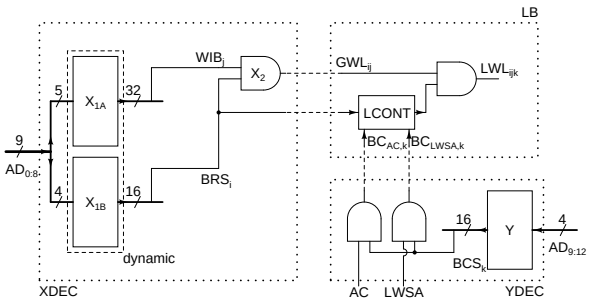


Fig. 3. The distributed decoder architecture. Decoding (select) and control (timing) signals are merged for efficiency.

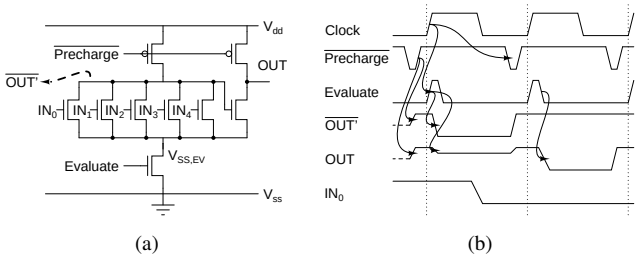


Fig. 4. A novel dynamic OR-gate adapted from [4], (a) circuit, (b) timing diagram.

B. Dynamic X-decoder implementation

The $9\,\mathrm{bit}$ to 512 GWL X-decoder consists of two stages. The first stage ($\mathrm{X_{1A}}$ en $\mathrm{X_{1B}}$) combines 5 address bits to 32 within-block signals (WIB) and 4 address bits to 16 BRS signals. It is implemented with dynamic logic to gain speed and to avoid the need to store the address bits. The second stage ($\mathrm{X_2}$), which combines BRS and WIB signals, is implemented with static logic to save energy. The dynamic logic gate used for the first stage [4] is shown in Fig. 4. Nodes $\overline{\mathrm{OUT}'}$ and OUT are dynamic. They are precharged when $\overline{\mathrm{Precharge}}$ is low. When $\mathrm{Evaluate}$ becomes high, $\overline{\mathrm{OUT}'}$ gets discharged, except when all inputs IN_i are low. OUT only gets discharged when $\overline{\mathrm{OUT}'}$ was not, making it an OR-gate. By construction, this gate is glitch free, again saving energy. By using a pulsed version of $\mathrm{Evaluate}$, this gate functions as a flipflop. This eleminates the delay of the flipflops normally used to store the address bits. The energy overhead caused by the use of dynamic logic is limited as these gates are only used in the first stage where the number of gates is still limited (i.e. 48 in this design).

As OUT could be discharged before $\overline{\mathrm{OUT}'}$, which would lead to failures, stability of this gate under mismatch needs to be confirmed. Fig. 5 shows the failure rate in the worst case condition (only one IN_i high) as function of the gate output load. When the load is larger than 6 minimal inverters, correct operation is ensured.

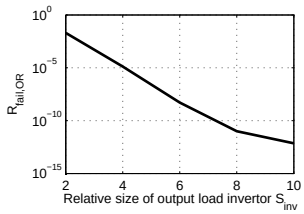


Fig. 5. Failure rate $R_{\mathrm{fail,OR}}$ of a 5-input dynamic OR-gate a.f.o. the load S_{inv}.

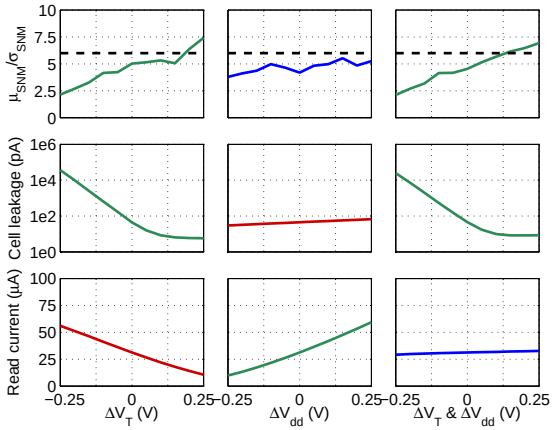


Fig. 6. Main properties of an SRAM cell a.f.o. changes in supply and threshold voltage. ($V_{\mathrm{dd,nom}} = 1.0\,\mathrm{V}$, $V_{\mathrm{T,nom}} = 0.25\,\mathrm{V}$) Increasing both threshold and supply voltage increases stability and reduces leakage without compromising read speed. The dashed line indicates the 10^{-9} failure rate.

IV. Cell Design

To save area, the SRAM cell of choice is still the classical 6T cell. In deep submicron technologies, local transistor mismatch caused by effects such as random dopant fluctuations and line edge roughness cannot be ignored. Fig. 6 shows the SNM (in numbers of σ's), the leakage power and the read current of a cell as a function of the supply voltage (V_{dd}) and the transistor threshold voltage (V_{T}). Increasing V_{T} strongly reduces leakage current and has a positive effect on read stability. However, read current decreases significantly. A higher V_{dd} increases stability and read current. When both V_{T} and V_{dd} increase, stability and leakage improve while read current does not change (the gate-source overdrive voltage of the current sinking transistors in the cell remains constant). This design uses high threshold voltage cell transistors and a cell supply voltage of $1.0\,\mathrm{V}$.

Cell speed (read current) and cell read stability can be improved further. A technique which accomplishes both is the use of local bitlines (LBL) with local read buffers [3]. By reducing the load which has to be discharged by the small memory cells, read time is reduced. Stability is improved as critically unstable cells will not have flipped before the LBL is discharged. This is only possible when the bitline load is small enough. To qualify this, a variation on the classic SNM stability metric called *transient* static noise margin (SNM_{tran}) is used [3]. Fig. 7 shows SNM_{tran} and the failure rate $R_{\mathrm{fail,cell}}$

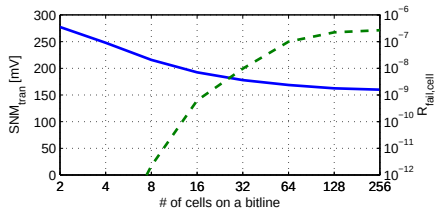


Fig. 7. SNM_{tran} (solid line) and $R_{\mathrm{fail,cell}}$ (dashed line) of an SRAM cell a.f.o. the number of cells on a bitline.

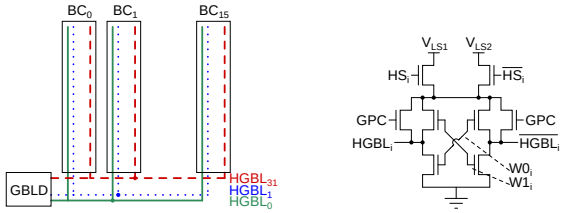


Fig. 8. (a) Granularity of dual swing voltages. Dual swing voltages are selected at HGBL level. A HGBL uses the same voltage for read and for write. Hence the full memory requires only 32 configuration bits HS_i. (b) Implementation of the GBL driver and precharge circuits (GBLD). $\mathrm{W0}_i$, $\mathrm{W1}_i$ and GPC are activated respectively when a zero or a one needs to be written or when the GBLs need to be precharged for read.

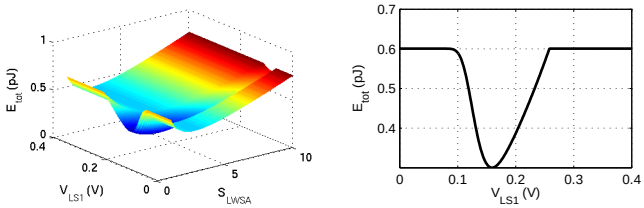
a.f.o. the number of cells on a local bitline (N_{cell}). Wire capacitance per cell is estimated to be $0.25\,\mathrm{fF}$.

V. Dual Swing Data Link

Sense amplifier (SA) design in the context of variability is a topic of ongoing research. Promising techniques to improve SA offset include SA tuning [5] and SA redundancy [6], [7]. Due to area considerations, these large or complex SAs cannot be used in the local blocks. This design proposes a different approach: dual swing data links. When a SA fails, this can be solved by increasing the signal swing at its inputs. However, this increases power quadratically. We propose the use of two different precharge voltages on the GBL. First, a low precharge voltage (V_{LS1}) will be used on most lines, ensuring low power. Second, a higher precharge voltage V_{LS2} is used on lines which contain failing LWSAs. Fig. 8(a) shows the granularity at which the dual voltages are applied. To avoid the need of full swing HGBL or of SAs between the HGBL and the VGBL, each low swing voltage is shared between a HGBL pair and the 16 VGBL pairs connected to it. Fig. 8(b) shows the implementation of the global bitline drivers (GBLD) which determine the voltages on the global bitlines for both read (precharge) and write (data).

If the offset of a LWSA follows a Gaussian distribution, the total energy E_{tot} consumed by the global bitlines can be calculated as follows. $R_{\mathrm{fail,GBL}}$ is the probability that at least one of the LWSAs on a single GBL fails when the low precharge voltage V_{LS1} is used. $C_{\mathrm{GBL,tot}}$ is the total capacitance of a global bitline. N_{GBL} is the number of global bitline pairs. E_{LWSA} is the switching energy of the LWSAs.

$$E_{\mathrm{tot}} = C_{\mathrm{GBL,tot}} \cdot N_{\mathrm{GBL}} \cdot (1 - R_{\mathrm{fail,GBL}}) \cdot V_{\mathrm{LS1}}^2 +$$



(a) (b)

Fig. 9. (a) E_{tot} a.f.o. V_{LS1} and S_{LWSA}, i.e. the LWSA size, for a die yield of 99.9 %, (b) Cross section at $S_{\mathrm{LWSA}} = 1$.

$$C_{\mathrm{GBL,tot}} \cdot N_{\mathrm{GBL}} \cdot R_{\mathrm{fail,GBL}} \cdot V_{\mathrm{LS2}}^2 + E_{\mathrm{LWSA}}$$

$R_{\mathrm{fail,GBL}}$ is a function of the lower precharge voltage V_{LS1} and the number of LWSAs on a single GBL. V_{deadzone} is the remaining voltage on the discharged bitline when the LWSA is triggered. It depends on the discharging time and on the threshold voltage of the discharging transistors. $R_{\mathrm{fail,LWSA}}$ is the probability of a single LWSA failing. N_{LWSA} is the number of LWSAs on a GBL.

$$R_{\mathrm{fail,LWSA}} = \mathrm{erfc}\left(\frac{V_{\mathrm{LS1}} - V_{\mathrm{deadzone}}}{\sqrt{2} \cdot \sigma_{\mathrm{offset,LWSA}}}\right)$$

$$R_{\mathrm{fail,GBL}} = 1 - \left(1 - R_{\mathrm{fail,LWSA}}\right)^{N_{\mathrm{LWSA}}}$$

The total energy was calculated for different sizes of the LWSA and different values of V_{LS1}. Fig. 9(a) shows the results. As LWSAs need to be as small as possible, Fig. 9(b) shows a cross section for LWSA size $S_{\mathrm{LWSA}} = 1$. V_{LS2} can be calculated from the required yield. For this design, it is 260 mV. A value for V_{LS1} which is too low does not decrease E_{tot} as no GBL will be able to use it. Obviously, when it is too high energy savings will drop again. In this design, the optimal global bitline energy E_{tot} for minimal sized LWSAs is 0.3 pJ for a V_{LS1} of 160 mV compared to 0.6 pJ for a classic single swing system.

Another advantage of the use of dual swing bitlines is that it can be used to compensate for a number of read failures such as GSA failure or cells that are too slow.

VI. Measurement results

The design was fabricated in a 65 nm low standby power CMOS technology. Fig. 10 shows a microphotograph of the die. Total memory area is 0.48 mm^2. Measurements show that the design is functional until 850 MHz. E_{write} is 4.6 pJ/access and E_{read} is 3.9 pJ/access resulting in an energy consumption of 4.3 pJ/access for a 50 % write pattern. Leakage is 25.2 µW. Table II shows a comparison to other recently published 65 nm memories with comparable size. FOM_{act} is the active energy per accessed bit. FOM_{leak} is the leakage power per bit in the memory. As can be seen, this memory reaches a much higher speed for less active energy compared to the state-of-the-art. Only memories with a speed which is orders of magnitude slower reach better leakage numbers.

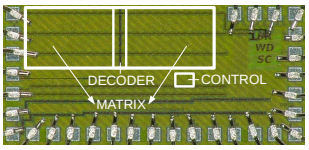


Fig. 10. The die microphotograph.

TABLE II
Comparison to state-of-the-art.
All designs are implemented in 65 nm CMOS.

Design	V_{dd} [V]	Size [kbit]	Speed [MHz]	E/acc [pJ]	FOM_{act} [fJ/bit]	P_{leak} [µW]	FOM_{leak} [pW/bit]
This	**1.00**	**256**	**850**	**4.3**	**133**	**25.2**	**96**
[7]	0.35	256	0.03	31	242	2.2	8.4
[8]	0.90	256	127	35	137	N/A	N/A
[9]	1.20	64	200	23	180	55	840

VII. Conclusion

A state-of-the-art 256 kbit memory has been designed, implemented and measured. Local word and bitlines ensure low power while dynamic stability and dual low swing lines ensure robustness. A dynamic first decoding stage results in a speed increase at a very moderate energy cost. Measured speed is 850 MHz for an energy per access of 4.3 pJ. Leakage power is only 25.2 µW due to the use of high threshold voltage cells.

References

[1] K. Masselos, F. Catthoor, C. Goutis, and H. Deman, "A systematic methodology for the application of data transfer and storage optimizing code transformations for power consumption and execution time reduction in realizations of multimedia algorithms on programmable processors," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 10, no. 4, pp. 515–518, Aug. 2002.

[2] M. De Nil, L. Yseboodt, F. Bouwens, J. Hulzink, M. Berekovic, J. Huisken, and J. van Meerbergen, "Ultra low power ASIP design for wireless sensor nodes," in *Electronics, Circuits and Systems, 2007. ICECS 2007. 14th IEEE International Conference on*, 2007, pp. 1352–1355.

[3] S. Cosemans, W. Dehaene, and F. Catthoor, "A low-power embedded SRAM for wireless applications," *Solid-State Circuits, IEEE Journal of*, vol. 42, no. 7, pp. 1607–1617, 2007.

[4] H. Nambu, K. Kanetani, K. Yamasaki, K. Higeta, M. Usami, T. Kusunoki, K. Yamaguchi, and N. Homma, "A 1.8 ns access, 550 MHz 4.5 Mb CMOS SRAM," in *Solid-State Circuits Conference, 1998. Digest of Technical Papers. 1998 IEEE International*, Feb. 1998, pp. 360–361, 464.

[5] S. Cosemans, W. Dehaene, and F. Catthoor, "A 3.6pJ/access 480MHz, 128Kbit on-chip SRAM with 850MHz boost mode in 90nm CMOS with tunable sense amplifiers to cope with variability," in *Solid-State Circuits Conference, 2008. ESSCIRC 2008. 34th European*, 2008, pp. 278–281.

[6] V. Sharma, S. Cosemans, M. Ashouei, J. Huisken, F. Catthoor, and W. Dehaene, "A 4.4pJ/access 80MHz, 2K word x 64b memory with write masking feature and variability resilient multi-sized sense amplifier redundancy for wireless sensor nodes applications," in *ESSCIRC, 2010 Proceedings of the*, 2010, pp. 358–361.

[7] N. Verma and A. Chandrakasan, "A 256 kb 65 nm 8T subthreshold SRAM employing sense-amplifier redundancy," *Solid-State Circuits, IEEE Journal of*, vol. 43, no. 1, pp. 141–149, 2008.

[8] K. Kushida, A. Suzuki, G. Fukano, A. Kawasumi, O. Hirabayashi, Y. Takeyama, T. Sasaki, A. Katayama, Y. Fujimura, and T. Yabe, "A 0.7 V single-supply SRAM with 0.495 um2 cell in 65 nm technology utilizing self-write-back sense amplifier and cascaded bit line scheme," *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 4, pp. 1192–1198, 2009.

[9] M. Sinangil, N. Verma, and A. Chandrakasan, "A reconfigurable 65nm SRAM achieving voltage scalability from 0.25-1.2V and performance scalability from 20kHz-200MHz," in *Solid-State Circuits Conference, 2008. ESSCIRC 2008. 34th European*, 2008, pp. 282–285.